John von Neumann Institute for Computing

Nearly-Deterministic Methods for Optimising Protein Geometry

G. Schenk, A. Torda

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 365-368, 2008.

http://www.fz-juelich.de/nic-series/volume40

Nearly-Deterministic Methods for Optimising Protein Geometry

Gundolf Schenk and Andrew Torda

Zentrum für Bioinformatik, Universität Hamburg, Bundesstraße 43, 20146 Hamburg, Germany
E-mail: {*schenk, torda*}*@zbh.uni-hamburg.de*

Protein structure prediction could be seen as either a challenge or an algorithmic playground. We are certainly interested in algorithmic improvements. Self consistent mean field methods (SCMF) have traditionally been used in areas such as wave function optimisation or protein side-chain placement. We have been trying to apply the ideas to find the most likely conformation for a protein. The philosophy relies on precalculated distributions of structural descriptors given a set of known properties (a protein's sequence). Starting with a sequence, which is decomposed into small overlapping fragments, the conformational space is described by a fixed number of weighted multivariate Gaussians (the known distributions). As the conformational bias, introduced by the sequence fragments, is local the weights of the Gaussians for all overlapping fragments can be optimised iteratively. Unlike molecular dynamics or Monte-Carlo simulations, the optimisation is done in probability space rather than on some initial structure. Therefore, we do not need to calculate energies as in standard SCMF. When the iteration converges sample structures are generated from the weighted Gaussians. The current results show that the procedure is able to find protein-like structures. We can also use this principle to predict protein sequences from structure.

1 Introduction

We are interested in self consistent mean field methods and the protein structure prediction problem. This also means formulating and building new force fields and treating also protein sequence optimisation. Our method has a probabilistic model of protein sequence-structure correlation and approaches self consistency within this framework.

Many methods have already been applied to ab initio protein structure prediction. All use some scoring schemes that are based on statistics and/or physics and chemistry. We want to avoid chemical detail as calculations become intractable and also coarse grain where one is usually dependent on preconceptions. Our approach is purely statistical with its own approximations, but little reliance on human preconceptions.

2 Methods

We have developed and successfully applied a scoring scheme to protein comparisons using sequence, structure or both[1,2]. It is based purely on Bayesian statistics and derived via a maximally parsimonious automatic classifier[3] from overlapping protein fragments. Each is described by 5-7 amino acid types and 10-14 dihedral angles from the backbone. The method assigns a fixed number of class weights (typically 150-300) to each fragment.

With this scoring framework we are able to generate protein structure samples in four steps (figure 1). First, the class weights matrix is build from the sequence. Then, the conformational space is narrowed down by iteratively updating the class weights of overlapping fragments. The local preferences are propagated, as the positions within a fragment

are correlated. This favours the consistent classes. From the conditional class weights sample structures can be generated. As a final step, steric clashes are removed and the models are collapsed by resampling random stretches. Unlike in standard SCMF, the method works without assuming the Boltzmann distribution at any stage.

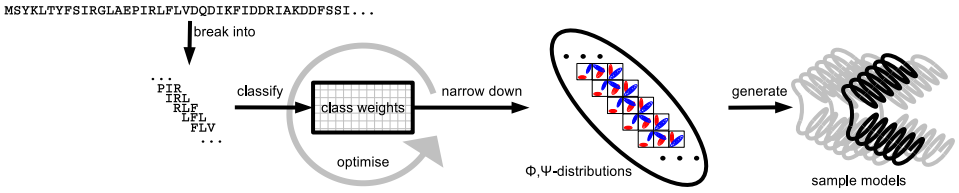

Figure 1. In the online phase a sequence is broken into overlapping fragments. Each fragment is classified leading to a probabilistic description of possible conformations. This can be used to generate sample structures.

The method is available as a web service:
`http://cardigan.zbh.uni-hamburg.de/~mahsch/schenk`

Given an amino acid sequence the server generates a huge number of samples and ranks them with their fragment probabilities,

$$\left(\prod_{i=1}^{N_{\text{fragments}}} \sum_{j=1}^{N_{\text{classes}}} w_{j|f_i^{\text{AA}}} \mathrm{p}_j \left(f_i^{\Phi,\Psi} \right) \right)^{\frac{1}{N_{\text{fragments}}}},$$

where N_{xx} is number of xx, $w_{j|f_i^{\text{AA}}}$ is the conditional weight of class j given the sequence fragment f_i^{AA} and $\mathrm{p}_j \left(f_i^{\Phi,\Psi} \right)$ denotes the conditional density of the structure fragment $f_i^{\Phi,\Psi}$ in class j.

3 Results

The Evaluation of 100000 samples of selected targets (figure 2) suggests that the target structure can be found among the generated models. α-helical targets seem easier than those containing β-strands, which is consistent with other methods. We also calculated the structure which corresponds to the distribution mean. However, we find it far from correct.

There are certain limits with the evaluation one should keep in mind when interpreting the results. The multivariate Gaussian model lacks to account for the periodic nature of dihedral angles and the use of idealised bonds lengths and angles during structure construction introduces a few Ångstrøms error.

Another application of the classification is the prediction of amino acid composition from structure. The regenerated sequences are about 20% identical to the original. So far, it is unknown whether these are bad sequences or alternative possibilities folding to similar structures.

Target Protein				Samples	Best	Highscore	Mean	Mean (opt.)	Samples (opt.)		FB5-HMM*		ROSETTA*	
PDB code	Length	α	β	<6Å [%]	RMSD [Å]	RMSD [Å]	RMSD [Å]	RMSD [Å]	<6Å [%]	RMSD [Å]	<6Å [%]	RMSD [Å]	<6Å [%]	RMSD [Å]
1FC2	43	2	0	3.510	4.1	6.2	6.5	5.5			9.593	2.7	95	3.3
1ENH	54	2	0	0.553	4.4	11.2	9.8	10.7	0.387	4.6	6.595	2.5	47	2.7
2GB1	56	1	4	0.002	5.5	11.3	9.4	9.9	0.001	5.8	0.037	4.9	0	6.3
2CRO	65	5	0	0.050	5.3	8.7	10.6	9.1	0.052	5.2	0.464	3.9	18	4.2
1CTF	68	3	3	0.003	5.6	11.0	12.7	11.0			0.009	5.4	6	5.3
4ICB	76	4	0	0.003	5.7	11.2	10.4	8.0	0.004	5.3	0.089	4.3	17	4.7

Figure 2. Evaluation of 100000 samples of selected targets. *Numbers taken from Ref. 4.

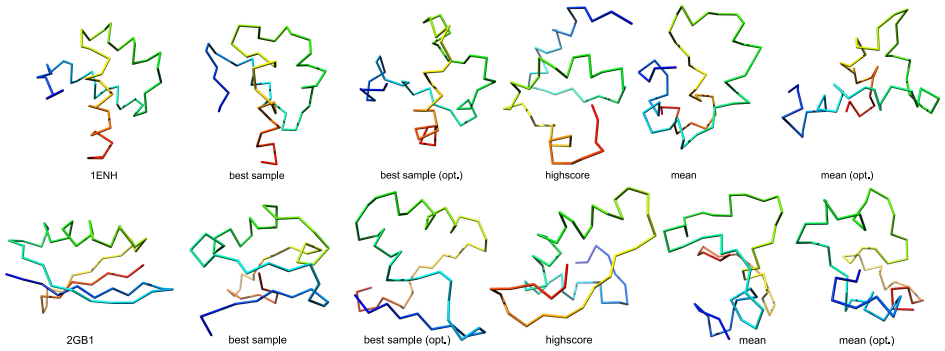

Figure 3. Two Examples from the evaluation.

4 Conclusions and Outlook

The method was tested with only a simple term to favour compaction. It produces the correct type of fold with secondary structure and loops in correct places. There is some limitation of coordinate reconstruction.

We are participating in the CASP8 competition. For this we are improving the server ranking. We are testing a combination with Monte Carlo optimisation methods. To improve the quality of the generated models we are incorporating solvation and long-range terms into our scoring functions. Finally, we are fastening our sampling method.

References

1. G. Schenk, T. Margraf and A. Torda, *Protein sequence and structure alignments within one framework*, Algorithms for Molecular Biology, 3:4, 2008.
2. T. Margraf and A. Torda, *Salami Server*, `http://www.zbh.uni-hamburg.de/salami`.
3. P. Cheeseman and J. Stutz, *Bayesian Classification (AutoClass): Theory and Results*, Advances in Knowledge Discovery and Data Mining, 1996.
4. T. Hamelryck, J. T. Kent and A. Krogh, *Sampling Realistic Protein Conformations Using Local Structural Bias*, PloS Comput Biol **2(9):e131**, 1121–1133, 2006.